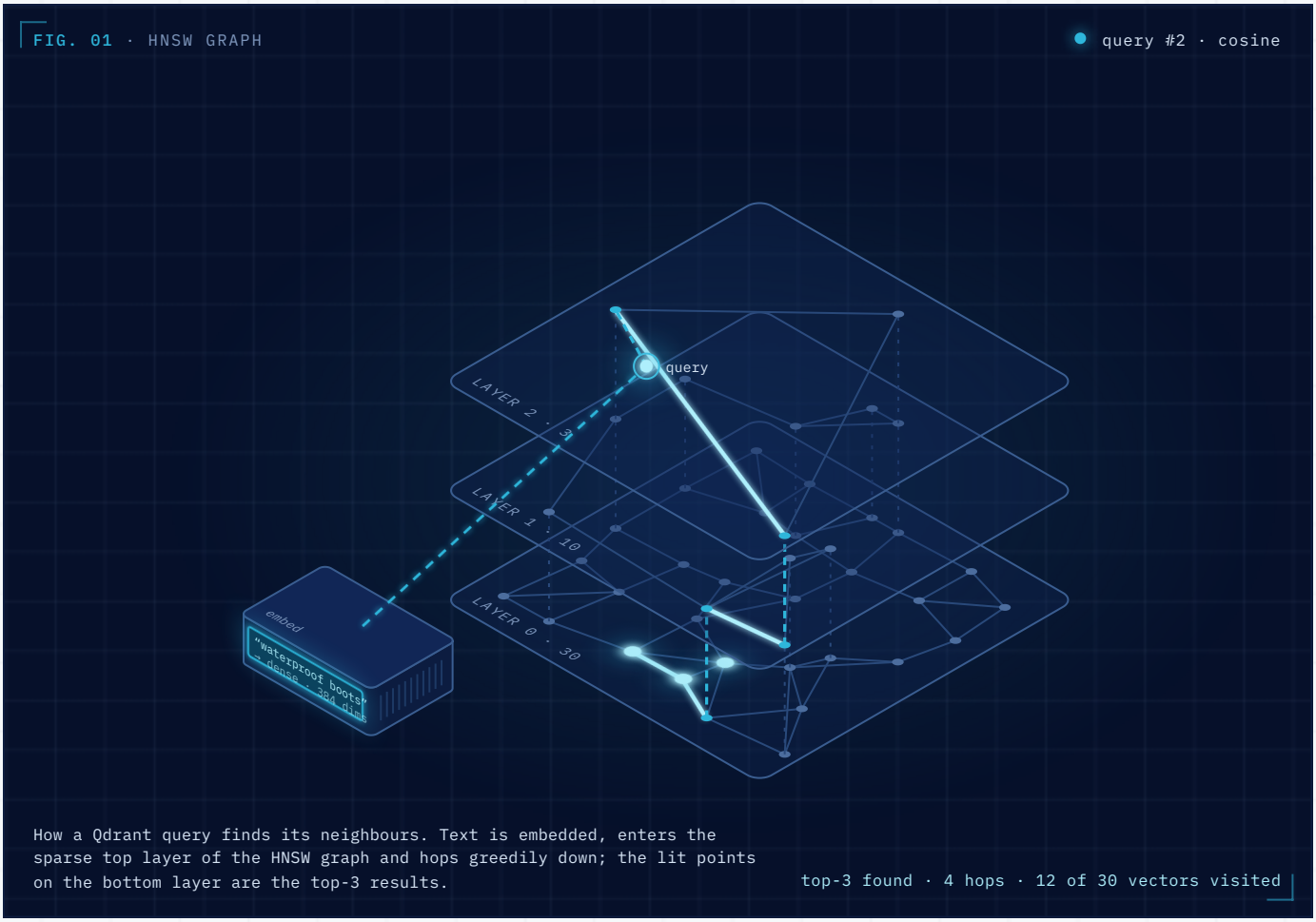
TRAINING & CONSULTING · VECTOR SEARCH

# Vector search that holds up *after* the demo.

**Consulting and training for teams building semantic search, RAG and recommendations on Qdrant.** We work on production Qdrant clusters every week — sizing them, tuning recall against latency, migrating embedding models and measuring relevance — and we teach what we see there.




## Weekly
Hands-on work with production Qdrant clusters

## 3 days
From the first embedding to a production cluster

## ≤ 15
People per class, with a cluster for each of them

## EN · TR
Delivered in English or Turkish, remote or on-site

### CONSULTING
**Get it right before it is big.** Sizing, quantization, filtering and scaling decisions made on your data, with the trade-offs measured.

### TRAINING
**Understand what the index is doing.** Embeddings, HNSW, hybrid search and operations — with labs on real datasets.

### SUPPORT
**Keep it fast as it grows.** Reviews before launches, version upgrades, model changes and incident help when you need it.

02 · CONSULTING

# Recall, latency and memory, *traded* on purpose.

Most vector search problems are choices made early and never measured: the wrong model, filters without indexes, quantization without rescoring, a shard layout that cannot grow. We find them, measure them and fix them with you.

01

### Architecture & sizing

Collections, named vectors, multitenancy, memory vs. on-disk storage and node count — for Cloud, Hybrid Cloud or self-hosted.

YOU GET

A sizing sheet (RAM and disk per million vectors) and a deployment plan.

02

### Search quality

Embedding model choice, hybrid dense + sparse retrieval with RRF, rerankers and filters, judged on your own queries.

YOU GET

A judged query set with recall and nDCG, before and after.

03

### Performance tuning

HNSW m and ef, payload indexes created before data, scalar, binary and product quantization with rescoring.

YOU GET

Latency vs. recall curves for your data and the settings to ship.

04

### Scaling & high availability

Sharding, replication factor, resharding, write consistency, snapshots and disaster recovery.

YOU GET

A tested failover and restore runbook.

05

### Model migration

Change or upgrade embedding models with zero downtime: parallel named vectors, backfill, shadow reads, alias switch.

YOU GET

A step-by-step migration plan with rollback.

06

### Production readiness

Prometheus and Grafana, optimizer health, memory growth, API keys and JWT access, version upgrades.

YOU GET

Dashboards, alerts and a reviewed upgrade path.

## SIGNALS IT IS TIME TO CALL

- Recall dropped after turning on quantization
- Filtered searches are slow or quietly miss results
- Memory keeps growing and nobody is sure why
- Segments stay unindexed and the collection is yellow
- You need to change the embedding model in production
- Thousands of tenants share one collection

## HOW AN ENGAGEMENT RUNS

01 **Scoping call** — 45 min · free
We hear the problem, agree goals and the numbers success is judged by.

02 **Assessment** — 1–2 weeks
Read-only access, diagnostics, logs and short interviews. Findings in writing.

03 **Fix & tune** — 2–4 weeks
Changes made with your team in small steps, each measured before and after.

04 **Hand-over** — ongoing
Runbooks, a focused workshop for your engineers and optional retained support.

FIELD EXPERIENCE: E-commerce search · RAG & assistants · Recommendations · Media · SaaS · Public sector · Banking & finance

03 · TRAINING

# From the first embedding to a *production* cluster.

Three days, one idea at a time, each followed by a lab on your own cluster. A shorter relevance workshop is available for teams that already run vector search and want to measure it properly.

PROGRAMME 01 · CORE

## Vector Databases with Qdrant

**3 days · 18 hours**

LEVEL Intermediate

PREREQ. Basic Linux, Docker and Python

- **Day 1 · Foundations.** Embeddings and distance metrics; vector vs. traditional databases; Qdrant architecture; Docker setup, Python client, REST and gRPC; a first RAG prototype
- **Day 2 · Search.** Batch upload, payloads and nested fields, named vectors; filters (must, should, must_not, range, geo); hybrid search, score thresholds, pagination
- **Day 2 · Index.** HNSW parameters, scalar and product quantization, indexing strategies, sharding and replication; Qdrant or Elasticsearch — when to use which
- **Labs.** An e-commerce product search and a semantic search engine, built end to end
- **Day 3 · Production.** API keys and JWT, snapshots and backups, monitoring and logging, clusters, resource limits, horizontal scaling, replicas, load balancing and disaster recovery
- **Day 3 · Practice.** Choosing an embedding model, filtering strategies, error handling, retries and rate limiting

PROGRAMME 02 · WORKSHOP

## Measuring Vector Search Relevance

**1 day · 6 hours**

LEVEL Intermediate – advanced

PREREQ. Programme 01 or equivalent

- Building a judged query set quickly, with AI-assisted drafts and human review
- Recall@k, precision, nDCG and MRR — what each one hides
- Exact vs. approximate search: measuring what HNSW misses
- Dense, sparse (BM25, SPLADE) and hybrid with RRF and DBSF
- Filters, payload indexes and the cost of creating them late
- Quantization with rescoring: memory saved vs. recall lost
- Rerankers and when they pay for themselves
- A weekly scorecard your team can keep running

### WHO IT IS FOR

- ◇ Engineers building search, RAG or recommendations
- ◇ Architects comparing Qdrant, Elasticsearch, pgvector

### WHAT YOUR TEAM LEAVES WITH

- Hybrid search on their own data, tuned on purpose
- A production checklist they can apply on day one

**DELIVERY**
Remote or on-site; daytime, evening or weekend sessions

**LABS**
A cloud lab for every participant, on realistic data

**CLASS SIZE**
Up to 15 people, so everyone gets hands-on time

**LANGUAGE**
English or Turkish, with all materials provided

**TAILORING**
Content, length and examples fitted to your stack

NEXT STEP

## Start with a 45-minute scoping call.

Tell us what is slow, wrong or about to grow. We reply within one working day with a suggested next step — consulting, training or both.

info@scaleyazilim.com

scaleyazilim.com/services/qdrant

Ankara · Türkiye — remote or on-site, worldwide